

Modelling the problem

1.1 Characteristics of the problem

...

1.2 Bounds on variables known a priori

1.2.1 Lower bound on L_{\max}

A lower bound on L_{\max} can be found using the lower bound on the completion date of each bucket q , where a bucket is defined as the set of batches with due date δ_q :

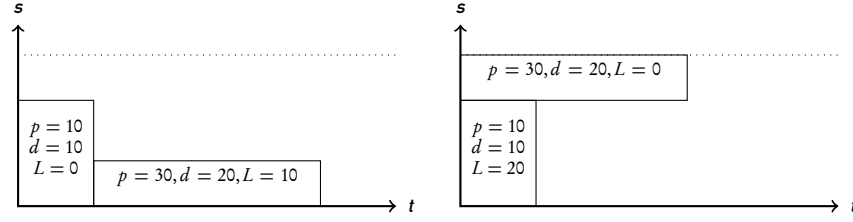
$$L_{\max} \geq C_{\max,q} - \delta_q \quad \forall q \quad (1.1)$$

This works because the buckets up to bucket q are guaranteed to contain all jobs with due dates $d \leq \delta_q$, since the batches within the buckets are ordered by earliest due date (EDD) in the optimal solution. The buckets up to bucket q will likely also contain some later ($d > \delta_q$) jobs in the optimal solution but this does not affect the validity of the lower bound.

Now we need to find $C_{\max,q}$, or at least a lower bound on it, in polynomial time. The simplest approach simply considers the “total elastic area”, i.e. the sum of all $s_j p_j$ products:

$$C_{\max,q} \geq \left\lceil \frac{1}{b} \sum_j s_j p_j \right\rceil \quad \forall q, \forall \{j \in J | d_j \leq \delta_q\} \quad (1.2)$$

A better lower bound on $C_{\max,q}$ would be given by a preemptive-cumulative schedule. Unfortunately, minimizing C_{\max} for such problems is equivalent to solving a standard bin-packing problem, which requires exponential time.

Figure 1.1: Overzealous batch elimination can increase L_{\max}

1.2.2 Upper bound on L_{\max}

An upper bound on L_{\max} can be found by using a dispatch rule to find a feasible, if not optimal, schedule. A good approach could be the “best-fit” heuristic proposed in the original paper. **This has not been implemented yet.**

1.2.3 Bounding the number of batches n_k

Initially, the number of batches needed is assumed to be equal to the number of jobs: $n_k = n_j$. Reducing n_k by pre-computing the maximum number of batches needed shrinks the x_{jk} matrix, and prunes potential search branches in branch-and-bound decomposition approaches.

Unfortunately, we cannot make a general statement that optimal solutions never have more batches than other feasible solutions – a simple counterexample is shown in figure 1.1.¹

Starting out with a one-job-per-batch schedule sorted by EDD, we can explore all feasible batch configurations recursively. To generate any other feasible schedule (including the optimal solution), jobs j are rescheduled (“moved back”) from their original batch k_{origin} into a prior batch k_{earlier} . This eliminates k_{origin} and requires, of course, that k_{earlier} has sufficient capacity. If the job’s processing time p_j exceeds that of k_{earlier} , then the lateness of batches between k_{earlier} and k_{origin} will increase; the merit of such a move cannot be judged in prior, so it is an *unsafe* batch elimination. *Safe* eliminations, on the other hand, will never worsen L_{\max} , and only they can be considered when bounding n_k a priori.

Algorithm 1 outlines a recursive method to find an upper bound on n_k , recognizing safe batch eliminations only.

¹To be more precise, we cannot state that at least one optimal solution is in the subset of feasible solutions that uses the fewest number of batches – a dominance situation that could be exploited, were it true.

Algorithm 1 Recursive algorithm to find an upper bound on n_k

```

if no open jobs left then                                     ▷ if this is a “leaf node” in the recursion
    update  $UB(n_k)$ 
end if
find the combination of unsafe later jobs that fills up the capacity most, leaving us with capacity  $b_r$ 
find all combinations  $x$  of safe later jobs that fit into  $b_r$ 
repeat
     $x$  = next safe job combination
     $ignoreJobs \leftarrow x$                                 ▷ make the moves, let the next recursion level deal with the rest of the jobs
    spawn and run child node with  $J \setminus ignoreJobs$ 
     $ignoreJobs \leftarrow ignoreJobs \setminus x$ 
until all combinations have been explored
return

```

This algorithm evidently requires exponential time. A relaxed variant is a possible option: if only a single unsafe move *into* a batch is possible, no safe eliminations into that batch are considered at all and we skip to the next batch. This would greatly speed up the recursion but also significantly weaken the usefulness of the resulting upper bound.

In a branch-and-bound decomposition approach in which batches are modelled individually, an upper bound on n_k could be used to limit the depth of the search tree, or, in combination with a method to determine a lower bound on the remaining jobs’ n_k at every node, to actively prune the search tree during the search. The latter method, however, would also run in exponential time as it, again, would require knapsack-type reasoning unless we use a much less powerful relaxation.

1.3 MIP model

Malapert’s original MIP approach is given in Model 1.3.1.

The constraint 1.8 is the EDD constraint.

1.3.1 Grouping empty batches

The given formulation lacks a rule that ensures that no empty batch is followed by a non-empty batch. Empty batches have no processing time and a due date only bounded by d_{\max} , so they can be sequenced between non-empty batches without negatively affecting L_{\max} . Since, however, desirable schedules have no empty batches scattered throughout, we can easily reduce the search space by disallowing such arrangements. The idea is illustrated in Figure 1.2.

$$\sum_{k \in K} x_{jk} = 1 \quad \forall j \in J \quad (1.3)$$

$$\sum_{j \in J} s_j x_{jk} \leq b \quad \forall k \in K \quad (1.4)$$

$$p_j x_{jk} \leq P_k \quad \forall j \in J, \forall k \in K \quad (1.5)$$

$$C_{k-1} + P_k = C_k \quad \forall k \in K \quad (1.6)$$

$$(d_{\max} - d_j)(1 - x_{jk}) + d_j \geq D_k \quad \forall j \in J, \forall k \in K \quad (1.7)$$

$$D_{k-1} \leq D_k \quad \forall k \in K \quad (1.8)$$

$$C_k - D_k \leq L_{\max} \quad \forall k \in K \quad (1.9)$$

$$C_k \geq 0, P_k \geq 0 \text{ and } D_k \geq 0 \quad \forall k \in K \quad (1.10)$$

Model 1.3.1: Malapert's original MIP model

A mathematical formulation is

$$\sum_{j \in J} x_{j,k-1} = 0 \rightarrow \sum_{j \in J} x_{jk} = 0 \quad \forall k \in K. \quad (1.11)$$

To implement this, we can write constraints in terms of an additional set of binary variables, e_k , indicating whether a batch k is empty or not:

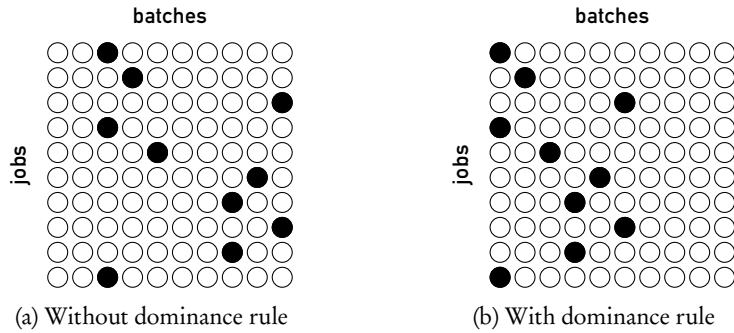


Figure 1.2: Dominance rule to eliminate empty batches followed by non-empty batches

$$e_k + \sum_{j \in J} x_{jk} \geq 1 \quad \forall k \in K, \quad (1.12)$$

$$n_j(e_k - 1) + \sum_{j \in J} x_{jk} \leq 0 \quad \forall k \in K. \quad (1.13)$$

Constraints (1.12) enforce $e_k = 1$ when the batch k is empty. Constraints (1.13) enforce $e_k = 0$ otherwise, since the sum term will never exceed n_j . The rule 1.11 can now be expressed as $e_{k-1} = 1 \rightarrow e_k = 1$, and implemented as follows:

$$e_k - e_{k-1} \geq 0 \quad \forall k \in K. \quad (1.14)$$

We can also prune any attempts to leave the first batch empty by adding a constraint $e_0 = 0$.

This, while it works just fine, actually takes longer than without it.

1.3.2 No postponing of jobs to later batches

Since the jobs are already sorted by non-decreasing due dates, it makes sense to explicitly instruct the solver never to attempt to push jobs into batches with a greater index than their own: even if every job had its own batch, it would be unreasonable to ever postpone a job to a later batch.

$$x_{jk} = 0 \quad \forall \{j \in J, k \in K | j > k\} \quad (1.15)$$

1.4 CP model

The mixed integer model can be turned into a constraint programming model with just a small number of modifications.

1.4.1 Bin-packing and cumulative constraints

This makes sure the jobs are distributed into the batches such that no batch exceeds the capacity b :

$$\text{pack}(J, K, b) \quad (1.16)$$

The cumulative constraint functions similarly, but instead of packing discrete bins, it enforces non-overlapping constraint on the temporal (IntervalVariable) J variables.

$$\text{cumul}(J, b) \quad (1.17)$$

1.4.2 Temporal constraints

These constraints are implemented using IntervalVar objects, which offer properties such as lengthOf or endOf.

$$P_k \geq \max_{j:B_j=k} p_j \quad \forall k \in K \quad (1.18)$$

$$D_k \leq \min_{j:B_j=k} d_j \quad \forall k \in K \quad (1.19)$$

$$C_k + P_{k+1} = C_{k+1} \quad \forall k \in K \quad (1.20)$$

$$L_{\max} \geq \max_k (C_k - D_k) \quad (1.21)$$

The first constraint ensures that each batch is as long as its longest job. The second constraint ensures that the earliest-due job sets the due date. The third constraint defines batch completion dates, and the fourth constraint defines L_{\max} .

The first two temporal constraints exploit the temporal features of IntervalVars.

1.4.3 Constraint on number of batches with length $P_k > p$

Since batches take on the processing time of their longest job, there is at least one batch with $P = \max_j p_j$:

$$\text{globalCardinality}(|P_k = \max_j p_j| = 1) \quad (1.22)$$

We can proceed to fill batches with jobs, ordered by non-increasing processing time, based on algorithm 2.

At the end of this algorithm, we can state:

$$\text{globalCardinality}(|P_{k-1,\min} > P_k \geq P_{k,\min}| \geq 1) \quad \forall k \in \{k_0, \dots, k_{LB(n_k)}\} \quad (1.23)$$

The algorithm sorts jobs by non-increasing p , and then fills batches job by job. If a job fits into a previous batch, it is assigned there. If a job fits into multiple

Algorithm 2 Generating lower bounds on batch lengths

```

 $J^* \leftarrow J$  ▷ initialize all jobs as unassigned jobs
 $n_k \leftarrow 1; S_k \leftarrow \{0\}; P_{k,\min} \leftarrow \{0\}$  ▷ Create one empty batch of size and length zero
sort  $J^*$  by processing time, non-increasing
repeat
   $j \leftarrow J^*.pop()$  ▷ select job for assignment, longest job first
  loop through all  $n_k$  existing batches  $k$ , first batch first
     $k_p \leftarrow \emptyset$  ▷ no feasible batch
     $c_{\min} = b$  ▷ currently known minimum remaining capacity
    if  $s_j < b - S_k$  and  $b - S_k < c_{\min}$  then
       $k_p \leftarrow k_p; c_{\min} \leftarrow b - S_k$ 
    end if
  end loop
  if  $|k_p| = 1$  then
     $S_{k_p} \leftarrow S_{k_p} + s_j$  ▷ assign job  $j$  to batch  $k_p$ 
  else
    if  $n_k < LB(n_k)$  then
       $n_k \leftarrow n_k + 1$  ▷ open new batch
       $S_{n_k} \leftarrow s_j; P_{n_k,\min} \leftarrow p_j$  ▷ assign  $s_j$  and  $p_j$  to the new batch
    else
      leave the loop now and end.
    end if
  end if
until  $J^*$  is empty

```

previous batches, it is assigned to the batch with the smallest remaining capacity. This is sometimes called *best-fit decreasing* rule, and works as follows: let J^* be the set of jobs sorted by p , then at least one batch will be as long as the longest job j_1^* . If the next n jobs fit into this batch, then there is at least one batch not shorter than j_{n+1}^* , and similarly for subsequent batches. Unfortunately, optimal solution may perform better than the packing heuristic in terms of “vertical” (s_j) bin packing, and may thus require fewer batches. We therefore need to find a lower bound $LB(n_k)$ on the number of batches, and we can only guarantee the first $LB(n_k)$ of the above constraints to hold in the optimal solution. Finding a true lower bound is a two-dimensional bin packing problem, which runs in exponential time ... so we have to come up with an even lower bound – right now I can only think of j_0 , the number of jobs ordered by decreasing s_j that can never fit into a batch together.

Furthermore, if all jobs have different processing times, all batches will have different processing times as well: $\text{alldifferent}(P_k)$. If m out of n_j jobs have different processing times, we can still enforce $\text{k_alldifferent}(P_k, m)$. Propagation rules for this constraint were given in [?]. I can’t find anything on this w.r.t. CP Optimizer,

so I may have to implement this myself ... time permitting.

1.4.4 Temporal constraints on a job's start date

Given any partial assignment of jobs and an open job j , we can reason that

- if the first batch with a due date later than the job is k , then the job cannot be part of a batch after k – this would result in a non-EDD sequence of batches.
- if the first batches up to $k - 1$ offer not enough capacity for j due to the given partial assignment, then the job cannot be part of a batch before k .

Since batches are *not* dynamically created like in Malapert's solution but fixed from the start, any partial assignment that fails due to these constraints cannot be part of an optimal solution.

This constraint is redundant with both the $(C_{k+1} \geq C_k)$ and packing constraints, but may help accelerate the propagation in some cases.

1.4.5 Grouping empty batches

Just like in the MIP model, we can force empty batches to the back and thus establish dominance of certain solutions. The implementation is much easier than in the MIP model:

$$\text{IfThen}(C_{k+2} > C_k, C_{k+1} > C_k) \quad \forall k \in \{k_1, \dots, k_{n_k-2}\} \quad (1.24)$$

1.4.6 No postponing of jobs to later batches

Just like in the MIP model, jobs should never go into a batch with an index greater than their own:

$$x_{jk} = 0 \quad \forall j, k : j > k \quad (1.25)$$

This is implemented as assignments $j \leq k$.

This constraint is analogous to the concept of finding an upper bound on n_k – essentially, it would be very helpful to find an upper bound on the latest batch for *every* job.

1.5 Decomposition approach

Instead of solving the entire problem using one model, we can solve the problem step by step, using the best techniques available for each subproblem. This approach is inspired by a method called *Benders' decomposition*.

1.5.1 Branch-and-bound by batch in chronological order

A basic version of this approach uses branch-and-bound to transverse the search tree. At each node on level ℓ , a single MIP and/or CP model is run to assign jobs to batch $k = \ell$. The remaining jobs are passed to the children nodes, which assign jobs to the next batch, and so on – until a solution, and thus a new upper bound on L_{\max} , is found. Several constraints are used to prune parts of the search tree that are known to offer only solutions worse than this upper bound. Figure 1.3 shows an example in which a MIP model is used to assign jobs to the batch at every node. Algorithm

Algorithm 3 MIP node class code overview

```

update currentAssignments                                ▶ this keeps track of where we are in the tree
if no jobs given to this node then                        ▶ if this is a leaf node, i.e. all jobs are assigned to batches
    calculate  $L_{\max, \text{current}}$  based on currentAssignments
    if  $L_{\max, \text{current}} < L_{\max, \text{incumbent}}$  then
         $L_{\max, \text{incumbent}} \leftarrow L_{\max, \text{current}}$ 
        bestAssignments  $\leftarrow$  currentAssignments
    end if
    return to parent node
end if
set up MIP model                                          ▶ as described below
repeat
     $x_j \leftarrow \text{model.solve}(x_j)$                     ▶ let model assign jobs to the batch
    spawn and run child node with all  $\{j | x_j = 0\}$       ▶ pass unassigned jobs to children
    add constraint to keep this solution from recurring    ▶ this happens once the child node returns
until model has no more solutions
return to parent node

```

3 outlines what happens at each node: the model finds the best jobs to assign to the batch according to some rule, lets the children handle the remaining jobs, and tries the next best solution once the first child has explored its subtree and backtracked.

Using MIP and cumulative packing after the batch

The first version of the branch-and-bound batch-by-batch decomposition uses a MIP model at each node to assign jobs to the respective batch. The remaining jobs are

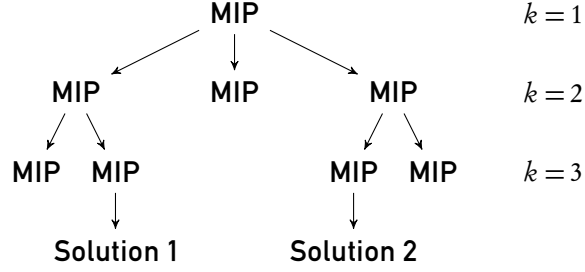


Figure 1.3: Batch-by-batch decomposition using MIP only

packed such as to minimize their L_{\max} , with a relaxation of the batching requirement, i.e., as if on a cumulative resource.

x_j	is 1 iff job j is assigned to the batch
u_{jt}	is 1 iff job j starts in time slot t
T_{jt}	is the set of all time slots occupied by job j if it ended at time t , that is $T_{jt} = \{t - p_j + 1, \dots, t\}$
v_k	is the start time of the batch at the given node in the search tree
$L_{\max, \text{incmb}}$	is the incumbent (known best) value of and thus an upper bound on L_{\max}
\mathcal{H}	is the set of all indexed time points $\{1, \dots, n_t\}$
w_j	is 1 iff job j is either longer than the batch ($p_j > P_k$) or already part of the batch ($x_j = 1$)

Table 1.1: Notation used in the decomposition model

Model 1.5.1 implements a time-indexed cumulative constraint on the non-batched jobs. Constraints (1.27) through (1.29) ensure that the batch stays below capacity, define the duration of the batch P_k and force at least one of the earliest-due jobs into the batch.

Constraints (1.30) express the interval relaxation used to ensure that no jobs exceed their latest allowable finish date given any batch assignment. Even jobs that are assigned to the batch have to fulfill this requirement.

Constraints (1.31) and (1.32) implement the cumulative nature of the post-batch assignments by ensuring that each job starts only once, and no jobs overlap on a given resource at any time (for the purposes of this model, the batch machine is considered divisible into b unary resources).

Constraints (1.33) again limit the possible end dates of a job, but unlike (1.30),

$$\text{Min. } L_{\max, \text{cumul}} \quad (1.26)$$

$$\text{s. t. } \sum_j s_j x_j \leq b \quad \forall j \in J \quad (1.27)$$

$$P_k \geq p_j x_j \quad \forall j \in J \quad (1.28)$$

$$\sum_j x_j \geq 1 \quad \forall \{j \in J \mid d_j = \min(d_j)\} \quad (1.29)$$

$$P_k + \frac{1}{b} \sum_i s_i p_i \leq d_j + L_{\max, \text{incmb}} - 1 - v_k \quad \forall j \in J, \forall \{i \in J \mid d_i \leq d_j\} \quad (1.30)$$

$$\sum_t u_{jt} = 1 \quad \forall j \in J \quad (1.31)$$

$$\sum_j \sum_{t' \in T_{jt}} u_{jt'} \leq b \quad \forall t \in \mathcal{H} \quad (1.32)$$

$$(v_k + t + p_j) u_{jt} \leq d_j + L_{\max, \text{incmb}} - 1 \quad \forall j \in J, \forall t \in \mathcal{H} \quad (1.33)$$

$$L_{\max, \text{cumul}} \geq (v_k + t + p_j) u_{jt} - d_j \quad \forall j \in J, \forall t \in \mathcal{H} \quad (1.34)$$

$$u_{j, t=1} = x_j \quad \forall j \in J \quad (1.35)$$

$$u_{it} \leq (1 - x_j) \quad \forall i, j \in J, \forall t \in \mathcal{H} \setminus \{2\} \quad (1.36)$$

$$b - \sum_j s_j x_j \leq (b w_j + 1) s_j \quad \forall j \in J \quad (1.37)$$

$$P_k + 2w_j n_t \geq p_j + n_t x_j \quad \forall j \in J \quad (1.38)$$

$$P_k - 2(1 - w_j) n_t \leq p_j + n_t x_j - 1 \quad \forall j \in J \quad (1.39)$$

Model 1.5.1: MIP model in batch-by-batch branch-and-bound

they use the time assignments on the cumulative resource to determine end dates. Constraints (1.34) define the value of $L_{\max, \text{cumul}}$, the maximum lateness of any job in the non-batched set.

Constraints (1.35) force batched jobs to start at $t = 0$, while (1.36) force *all* jobs to start either at $t = 0$ or after the last batched job ends.

Constraints (1.37) enforce a dominance rule: jobs must be assigned to the batch such that the remaining capacity, $b_r = b - \sum_j s_j x_j$, is less than the size s_j of the

smallest job from the set of non-batched jobs that are *not longer* than the current batch. That is, if there exists a non-batched job j with $p_j \leq P_k$ and $s_j \leq b_r$, then the current assignment of jobs is infeasible in the model. The reasoning goes as follows: given any feasible schedule, assume there is a batch k_a with b_r remaining capacity and a later batch k_b containing a job j such that $s_j \leq b_r$ and $p_j \leq P_{k_a}$. Then job j can always be moved to batch k_a without negatively affecting the quality of the solution: if the schedule was optimal, then moving j will not affect L_{\max} at all (otherwise, it was no optimal schedule); if the schedule was not optimal, then L_{\max} will stay constant (unless j was the longest job in k_b and L_{\max} occurred in or in a batch after k_b , in which case L_{\max} will be improved).

This rule is implemented by means of a binary variable w_j , which, as defined by constraints (1.37) and (1.38), is 1 iff $p_j > P_k \vee x_j = 1$. These are the cases in which a job j is *not* to be considered in (1.36), and so w_j is used to scale s_j to a value insignificantly large in the eyes of the constraint's less-than relation.

After a solution is found, a child node in the search tree has run the subtree and returned, a constraint of the form

$$\sum_j x_j + \sum_i (1 - x_i) \leq n_j - 1 \quad \forall \{j \in J | x_j = 1\}, \forall \{i \in J | x_i = 0\} \quad (1.40)$$

is added to the model before the solver is called again, to exclude the last solution from the set of feasible solutions.

Using CP and cumulative packing after the batch

This approach is equivalent, but we now use CP to select the batch assignments, again based on a minimized L_{\max} among the non-batched jobs.

1.5.2 Potential improvements

Improve the initial $L_{\max, \text{incumbent}}$ A better initial upper bound on L_{\max} can help prune some branches of the search tree from the outset. There are several dispatch rules (or maybe other heuristics?) that could be explored to do this better.

Improve $L_{\max, \text{incumbent}}$ during search It may be useful to use a heuristic like above to “complete the schedule” once a promising partial schedule has been generated. I have yet to identify situations where this is always helpful.

$$\text{Min. } L_{\max} \quad (1.41)$$

$$\text{s.t. } \text{IfThen}(x_j = 0, \text{startOf}(j) \geq P_k) \quad \forall j \in J \quad (1.42)$$

$$\text{IfThen}(\text{startOf}(j) \geq 2, x_j = 0) \quad \forall j \in J \quad (1.43)$$

$$\text{IfThen}(x_j = 1, \text{startOf}(j) = 0) \quad \forall j \in J \quad (1.44)$$

$$\text{IfThen}(\text{startOf}(j) = 1, x_j = 1) \quad \forall j \in J \quad (1.45)$$

$$P_k \geq x_j p_j \quad \forall j \in J \quad (1.46)$$

$$\text{endOf}(j) = d_j + L_{\max, \text{incmb}} - 1 \quad \forall j \in J \quad (1.47)$$

$$L_{\max} \geq \text{endOf}(j) - d_j \quad \forall j \in J \quad (1.48)$$

$$\text{IfThen}(p_j \leq P_k \wedge x_j = 0, b - \sum_{j \in J} s_j x_j \leq s_j) \quad \forall j \in J \quad (1.49)$$

$$P_k + \frac{1}{b} \sum_i s_i p_i \leq d_j + L_{\max, \text{incmb}} - 1 - v_k \quad \forall j \in J, \forall \{i \in J | d_i \leq d_j\} \quad (1.50)$$

$$\sum_j x_j \geq 1 \quad \forall \{j \in J | d_j = \min(d_j)\} \quad (1.51)$$

$$\text{cumul}(J, b) \quad (1.52)$$

Model 1.5.2: CP model in batch-by-batch branch-and-bound

1.6 Empirical results

Malapert uses benchmark data by Daste. For design purposes, I created my own set of randomized job lists, with $s_j, p_j \in [1, 20]$ and $d_j \in [1, 10n]$ where n is the number of jobs. Here is the Python code used:

```

1 import random
2
3 random.seed(None)
4
5 for i in range(11, 20) + range(10, 100, 10):
6     for j in range(1, 10):
7         with open("data_" + str(i) + "_" + str(j), "wb")
8             as csvfile:

```

```
9         csvwriter = csv.writer(csvfile, delimiter=",")
10        csvwriter.writerow(['s_j', 'p_j', 'd_j'])
11        for job in range(i):
12            csvwriter.writerow([random.randrange(1,20),
13                               random.randrange(1,20),
14                               random.randrange(1,5*i)])
15    csvfile.close()
```

CSV files can be read in with `IloCsvReader` and `IloCsvLine`.

1.6.1 Comparison of CPU time used by various models

Figure 1.4 shows the CPU time used by the models. Ten sample job sets are used per number of jobs, represented as $[\{10.0, \dots, 10.9\}, \{11.0, \dots, 11.9\}, \dots]$ on the horizontal axis. The models were run on an i7 Q740 CPU, which has 2×4 cores. Cplex supports multithreading and as such, elapsed wall clock time is significantly lower in most cases.

The CP branch-and-bound model times out on most instances and is not shown here.

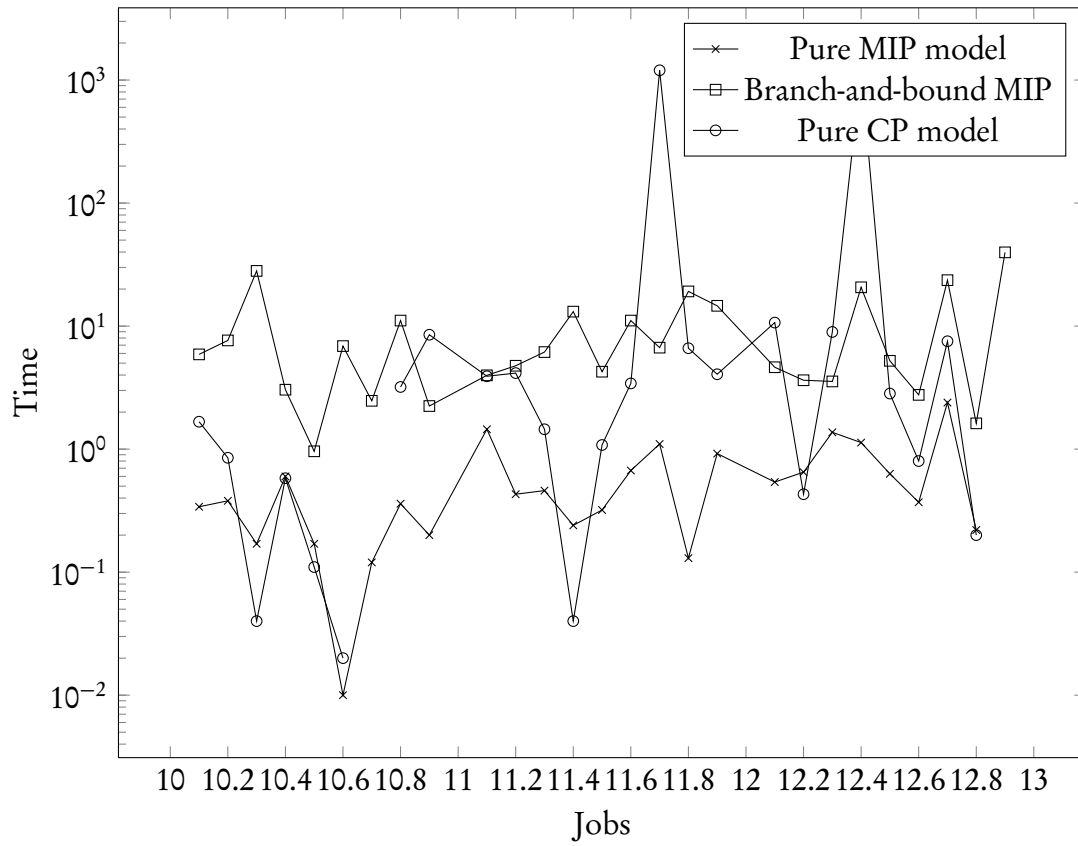


Figure 1.4: Comparison of CPU time used by different models.